Exploring design principles for data literacy activities to support children's inquiries from complex data

Journal Item

Exploring design principles for data literacy activities to support children's inquiries from complex data

Annika Wolff [1], Michel Wermelinger, Marian Petre

School of Computing and Communications, The Open University
Milton Keynes MK7 6AA, United Kingdom

Abstract

Data literacy is gaining importance as a general skill that all citizens should possess in an increasingly data-driven society. As such there is interest in how it can be taught in schools. However, the majority of teaching focuses on small, personally collected data which is easier for students to relate to. This does not give the students the breadth of experience they need for dealing with the larger, complex data that is collected at scale and used to drive the intelligent systems that people engage with during work and leisure time. Neither does it prepare them for future jobs, which increasingly require skills for critically querying and deriving insights from data.

This paper addresses this gap by trialling a method for teaching from complex data, collected through a smart city project. The main contribution is to show that existing data principles from the literature can be adapted to design data literacy activities that help pupils understand complex data collected by others and form interesting questions and hypotheses about it. It also demonstrates how smart city ideas and concepts can be brought to life in the classroom.

The Urban Data School study was carried out over two years in three primary and secondary schools in England, using smart city datasets. Three teachers took part, providing access to different age groups, subject areas, and class types. This resulted in four distinctive field studies, with 67 students aged between 10-14 years, each lasting a few weeks within the two year period. The studies provide evidence that when engaging with data that has not been personally collected, activities designed to give the experience of

[1]School of Engineering Science, LUT University, Lappeenranta, Finland

collecting the data can help in critiquing it.

Keywords: data literacy, human-data interaction, smart city, open data

1. Introduction

Society is increasingly driven by data. One example of its use is to inform business decisions, a process that is often referred to as business intelligence. With an increase in data available to businesses, there is a growing gap in the number of employees with the skills to make good use of it. In a policy briefing, Nesta explores this skills gap in detail and proposes ways to address it [1]. Amongst these is a proposal that highlights the importance of initiatives to teach data skills in school and to embed them into other subjects, improving the data literacy of school-leavers and their readiness for the future job market.

Business and employment needs are not the only drivers toward increasing data literacy. Presentation of online content is often decided based on analysis of what users having been clicking through or purchasing online, with the intention to influence the end-users' actions and decision-making. Examples include the recommendations made on shopping sites or entertainment services. Mortier et al. [2] argue that it is important to explore the issue of transparency of how users' data is collected and analysed and how to give increased agency to users who provide data so that they can themselves derive value from it. This is reliant on users having a level of data literacy that enables them to engage with their own data. Beyond this, a white paper of Bhargava et al. [3] highlights the importance of data literacy as an increasingly important skill for civic empowerment. Policy decisions and media reporting are increasingly justified with data, and people therefore need skills to assess critically the accuracy of what is presented to them as fact [4]. One final, yet important, reason for advocating data literacy is that citizens increasingly use data-driven smart technologies to make their lives more efficient, including smart meters, travel apps, or the currently popular 'sharing economy' apps through which people swap knowledge, goods and services. The increasing availability of open data is often mentioned as something that can support 'bottom up' citizen innovation, but this is predicated on citizens having appropriate skills to design around large, complex data sets. However, evidence provided by Janssen et al. [5] shows that this potential is not being reached, and that one of the key barriers is lower levels of data literacy amongst the general population.

To understand why this is the case, we turn our attention to what students are learning in school. Most of the examples mentioned above typically use large and complex data sets and require that people engage with data that they did not personally collect. In contrast, data sets traditionally used for teaching in schools tend to be smaller and are often collected by the students themselves. Research has shown that when analysing larger and more complex pre-existing data sets students may find it difficult to understand how the data were collected, which in turn makes it harder to interpret [6]. In general, skills learned on small data sets may not necessarily scale. This makes an argument for increasing the range of data used to teach data skills in school, which then raises the question how to achieve this in practice. At the same time, the work of Bowler and Acker [7] revealed that students' current understanding of data may be quite limited, for example they might understand the role of data in a scientific inquiry but not necessarily make the connections between their personal data and the different ways it may be used, or abused. Overall, this suggests that students may not be getting the broad data literacy learning that they need at an early age.

Despite its importance, there is currently little research that focuses on how to deliver data literacy teaching in the classroom, and in particular teaching that is based on analyzing more complex externally sourced data. This paper addresses this research gap by developing a method that draws on the existing approaches for teaching data literacy for smaller, personally collected data sets, and extends it to larger, externally sourced data. The main contribution is in the synthesis and reframing of existing principles to support the design of data literacy activities so that they can be adapted to this teaching context.

This paper reports on an exploratory two-year study in which these design principles were put to the test. Three teachers from three different UK schools took part in this initiative to integrate teaching data literacy skills into both primary and secondary school classrooms. The work described in this paper was conducted in the context of MK:Smart[2], a large smart city project in Milton Keynes. This project provided an opportunity to develop lesson plans and materials around some less typical data sets that were being collected as part of the project and at the same time to bring smart city

[2]http://www.mksmart.org

concepts into the classroom. The lesson plans were used in local schools. The approach taken was a user-centred ‘research through design’ [8, 9] approach that fit with the need to be flexible within each school engagement and in which each classroom engagement generated new knowledge. We discuss how the findings contribute to the following research questions:

- What factors influence students’ abilities to ask and answer questions from the presented data?
- What is the role of data interaction in facilitating the inquiry process?
- How does personally collecting a data set changes one’s perspective of it?

2. Background

There is no single agreed definition of data literacy and as a consequence, definitions can vary according to use. Wolff et al. [10] proposed the following definition to reflect the role of data for innovation:

“Data literacy is the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use of data. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to select, clean, analyse, visualise, critique and interpret data, as well as to communicate stories from data and to use data as part of a design process.” (p. 23)

Deahl [11] proposed that data literacy is: *“The ability to understand, find, collect, interpret, visualize, and support arguments using quantitative and qualitative data.”*

Hautea et al. [12] derived what they term *critical data literacies* using a bottom-up approach that observed young people’s interactions with data and how this helped them to articulate concerns about privacy and their scepticism around data accuracy, for example when they spotted inconsistencies in the data presented.

Despite this diversity of focus, there is a growing convergence on the idea that data literacy is more than simply learning a set of technical skills, such as how to read bar graphs [13, 14], work with maps [15] or use data for prediction [16]. While these are essential skills and worthy of study, other initiatives have taken a broader view of what is data literacy and how to develop it, especially within formal school education.

Among these approaches, several have focused on supporting data-driven inquiry. These include the work of Lee, Drak and Thayne [17] who used quantified self data to engage students with familiar personal data and then prompted them to drive their own inquiries from the data. The Local Ground project [18, 19] developed a geo-spatial data collection tool that students could use in geo-spatial data-driven inquiries. Dasgupta and Hill [20] supported children to drive their own inquiries from data and to create their own visualisations, using the Scratch programming environment, which many children are already using in school for programming. However, certain aspects of the inquiry process are found to be problematic, in particular how to link questions and data [21, 22].

Complementary to this, other approaches put the focus on the ability to use data for civic empowerment. These include the City Digits project [23] that aimed to teach data literacy skills to school children by encouraging them to investigate social issues in a local, urban context. Also, the Data Murals project [24] brought together a community to build an artwork that reflected their data explorations with data from and about their neighbourhood. Anslow, Brosz and Maurer [25] explore the potential of datathons for building data literacy, which bring together students and members of the community to solve problems.

Also gaining traction is a STEAM based approach. For example, D'Ignazio [26] focuses on approaches that support non-experts to learn important skills for framing problems around complex data through creative, rather than technical, activities.

Underpinning these, a number of principles to support data literacy learning have been proposed. These include the principles of data informed learning by Maybee and Zilinski [27] which propose that:

1. New ways of using data must build on students' prior experience.
2. Learning to use data should occur at the same time as learning about a disciplinary subject.
3. Learning should result in students becoming aware of new ways of using data as well as developing new understandings of the subject being studied.

Srikant and Aggarwal [16] proposed and tested these principles:

1. Use a full data cycle.

2. Make the data set relatable (e.g. about themselves).
3. Avoid pre-built data sets, but get students to do the task of data collection and entry themselves.
4. Reduce problem complexity (for example, if teaching predictive models, use only 2 categories).

Taking a slightly different approach, Bhargava and D'Ignazio [28] propose a set of design principles to use while developing tools to support data literacy learners, suggesting tools should be:

1. Focused, to do one thing well.
2. Guided, to help get the learner started.
3. Inviting, to appeal to the learner, maybe using data on a relevant or meaningful topic to the learner.
4. Expandable, offering paths to deeper learning.

The data literacy initiatives described have one thing in common, in that they focus on the use of data that is collected by the students themselves. As discussed, while clearly an essential skill, this does not necessarily translate to skills for dealing with externally sourced data [6]. Similarly, none of the data literacy design principles address this need, in fact the principles of [16] actively steer away from this, suggesting the students only engage with personal data. We instead propose to harness these same principles to *help* students engage with large, external data sets, through a small adaptation to a principle related to *personal data collection*. These principles are described in the following section. At the same time, there is little discussion in the literature of how such principles can be applied in practice, or how tools have been designed using principles for tool development described by [28]. We therefore show how these principles have been used to guide the co-creation of a set of lesson plans and the design of new tools that *complement* them, and then we explore how they are used in real classroom settings.

3. Data Literacy Activity Design Principles

We propose the following set of principles to support the design of activities for teaching data literacy, which synthesises the existing principles found in the literature. The main contribution is in the adaptation of a personal data collection principle (P6) to show how personal data collection can be used to complement interpretation of existing data, rather than to be used *instead* of it:

P1 Inquiry Principle: Follow an inquiry process to scaffold the data analysis. Lead the students first in a guided inquiry, from which follows an open inquiry when students are more familiar with the data and the approach. [16, 28].

P2 Expansion Principle: Start from a representative snapshot of a small part of the data set and expand out, rather than starting with the full, large data set and focusing in. This aims to help students' more easily relate questions to data [22] and to be expandable and offer paths to deeper learning [28]. It aims to provide students the opportunity to orient themselves within the data, before navigating across it, e.g., through time and/or space and/or some other dimension of the data.

P3 Context Principle: Teach in a context the student understands, using data that is from their own environment, either local to them, or else relating to them in some other way [27, 16, 28].

P4 Foundational competences principle: Focus on developing foundational competencies rather than practical skills, for example how to ask 'good' scientific questions from data [21, 22].

P5 STEAM principle: Take a STEAM approach by working collaboratively on creative activities alongside practical ones [26, 24].

P6 Personal Data Collection Principle: Students should engage with data they have collected themselves. When students are analysing an external data set, they should be given additional activities that support them in understanding what it is like to collect that type of data. This is to support them in contextualising and interpreting the data external data, which according to [6] they may otherwise struggle with.

The remainder of the paper describes how these principles have been used in practice to guide creation of lesson plans based around data collected within a smart city project. We focus particularly on evaluating the use of principle P6.

4. Iterative Design of Lesson Plans

The overall methodology can be categorised as *research through design*. This is a method in which design practice is applied to the creation of artefacts as a way of exploring solutions to problems, especially 'wicked problems' [8, 9]. In research through design, new knowledge is constructed by undertaking activities associated with design, such as iteratively creating and testing prototypes to understand and solve a problem and to act as a focal point for discussion by making interactions observable. This approach is fairly similar to that taken by data literacy initiatives, such as City Digits [23] and Data Murals, [24], though they are not necessarily framed that way. In our case, the research through design process was focused around the interpretation and use of the activity design principles to create lesson plans to teach data literacy skills and support interaction with smart city data and what we could learn by putting these into practice and through the iterative improvements to lesson plans over time. The relation between the design decisions and the design principles are highlighted throughout the text describing the lesson plans.

We adopted a user-centred iterative design approach with a small group of teachers. There were a number of stages: scoping; identifying potential data sets; drafting lesson outlines; creating an initial set of activities and lesson plans; introducing technologies. Each stage is described in turn.

Scoping: This first stage, which aimed to set boundaries on the types of activities that could be proposed, occurred prior to any engagement with schools. In this stage the decision was made to a) build activities that could be deployed using standard classroom equipment, technologies or software (e.g., iPads, desktop computers, web browsers) and b) build lesson plans from existing data sets, rather than being dependent on capture of data by students, e.g., through sensor technologies. This was in order to keep the initial focus on how to design learning experiences with these external data sets.

Identifying data sets: The second stage involved identifying a number of data sets that were available and could potentially be used for teaching. This resulted in a pack showing representative 'snapshot' visualisations of a small part of a number of data sets with some generalised lesson outlines that were broadly speaking agnostic of any particular teaching approach (e.g., inquiry-based, collaborative learning). These lesson outlines identified the types of questions that could be answered by the data, but did not propose any activities or constitute a lesson plan. They were intended to help teachers to understand the data, as it would be unfamiliar to them, and to act as a starting point for discussions. The chosen data sets were all related to the topic of *renewable energy*. They included smart meter data and data on solar energy potential for a number of houses in the city. They were at the time being used within smart city research into load shifting (trying to change typical patterns of energy use to times when overall demand for energy is lower) and in identifying new opportunities for solar installations or community energy solutions.

Lesson outlines: The third stage involved teachers from two schools, one primary mathematics teacher and one secondary science teacher, who had expressed an interest in using data from the smart city project in their classrooms. Each was invited to discuss the data sets and lesson outlines and how they could be formed into lesson plans. The possible use of an inquiry-based approach for teaching was also discussed. The teachers confirmed that these were not typical data sets used in teaching and were keen that students would get some experience in handling these different types of data. While the teachers came from different subject areas, the topic 'data inquiry' was seen to fit quite well in either mathematics or science, and ultimately the subject area did not play a big part in shaping the lesson plans. The secondary school science teacher was very familiar with an inquiry approach, as used in science, and was keen that this would be the approach used with the data.

Teaching activities and lesson plans: Through these discussions, the initial set of teaching activities and lesson plans was created, based on the principles P1-P6 described earlier. Tasks were adapted for each specific school context, based on the recommendations of the class teacher, so that the experience would align with what the students had been learning and be suited to their overall abilities. This allowed us to gain a better understanding of what the overall differences might be between schools and age groups, but ruled out a controlled approach to evaluation, across different school settings. These lesson plans are described in the next section.

Introducing technologies: The first trials were conducted using paper materials. Later trials introduced technologies to support interaction with the data, being focused on only simple functionality [28] to support key aspects of the task (as identified through first trials) and following the *expansion principle* (P2).

5. Lesson Plans

For each lesson plan, we describe: a) the overall aims of the lesson and the data set on which it was based, whether it was an existing data set or collected by the students for the purpose of contextualising one of the data sets; b) the activities undertaken with the data and how they were related to the design principles; c) the intended outcomes. The activities were used in various configurations across four separate field trials. The configuration was decided based on several meetings with the teacher. It should be noted that while there was never any need to adapt materials based on the classroom subject, the introduction that was given to the class prior to starting activities was different in each case, based on students' prior knowledge. These general introductions are not discussed further in this paper. Some other lesson plan variations were necessary due to the age of the students and also based on developments that happened in technology during the period of the project. These variations and their reason are indicated.

5.1. Lesson Plan 1 (LP1): Smart Meter Energy Data

The aim of this lesson was to show, through data, how energy consumption and generation from solar panels did not always match if people were not typically at home during the day when solar energy was being produced. This lesson used smart meter data from approximately 70 houses. For each property, students had access to data about: a) whole house consumption; b) individual appliance consumption; c) generation of solar energy. The example (figure 1) shows whole house consumption for one day in March. This data was anonymised, but it came from the same city that the students inhabited and this was conveyed to students to help them to contextualise the data (P3).

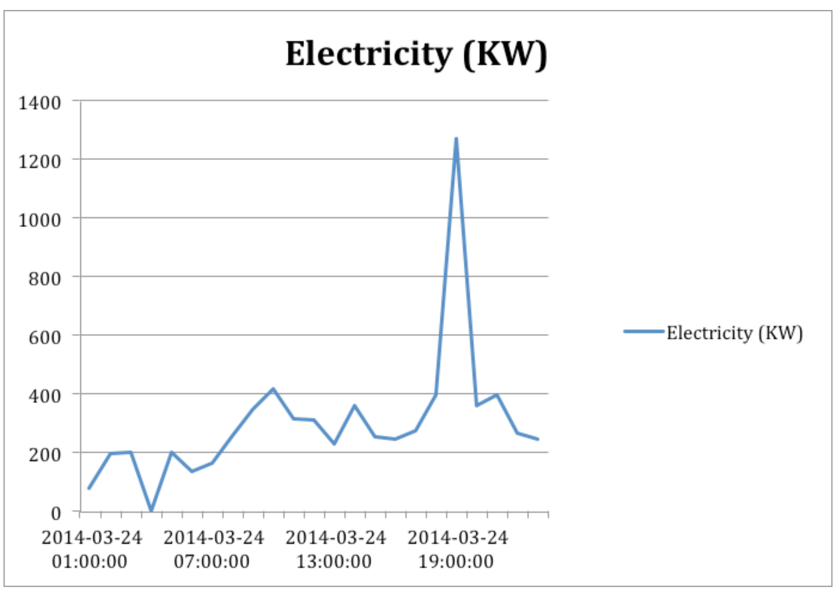

Figure 1: Smart meter data showing whole house consumption in one day

5.1.1. LP1 Activities

Students followed an inquiry process, based on posing questions from the data set (P1). The guided inquiry stage started with a snapshot of data (P2), as in figure 1, and some questions to answer from it. These asked when was most or least energy used and also prompted students to tell a story about the people living in that property, based on how they were using energy. Students worked in groups on all activities.

After familiarisation with the data, the next stage prompted students to explore the wider data set (P2), for example, answering questions about whether all houses showed the same pattern, or if the patterns varied at different times of year. There were variations in how this stage was delivered, which were tailored based on the age of the students and the development of technologies over the course of the project. The variations were as follows.

Guided: Students were guided using existing questions. This was used with younger students.

Guided, then Open: After the guided inquiry, students asked and answered their own questions. This had two stages, a brainstorming stage where students posed question and discussed them as a class, then a refinement stage, where they chose just one or two questions to follow up from the data (P4). This was used with older students.

No technology: Students worked from paper. Data was curated, either into further snapshots (guided activities) or based on the refinement stage, raw data was curated for students to explore one week later (open activities).

With technology: Students could ask and answer questions rapidly through the data browser (open activities). The data browser supported the selection of different houses. It followed approximately the design shown in the Balsamiq mockup in figure 2, with the exception that to configure the interface to view different houses required to first submit the house numbers and then select the rest of the attributes (time period, data).

5.1.2. LP1 Outcomes

The intended outcomes were that students would be able to use the data to identify common patterns in energy consumption and to see how these differ by day (e.g., weekday/weekend), household or time of year.

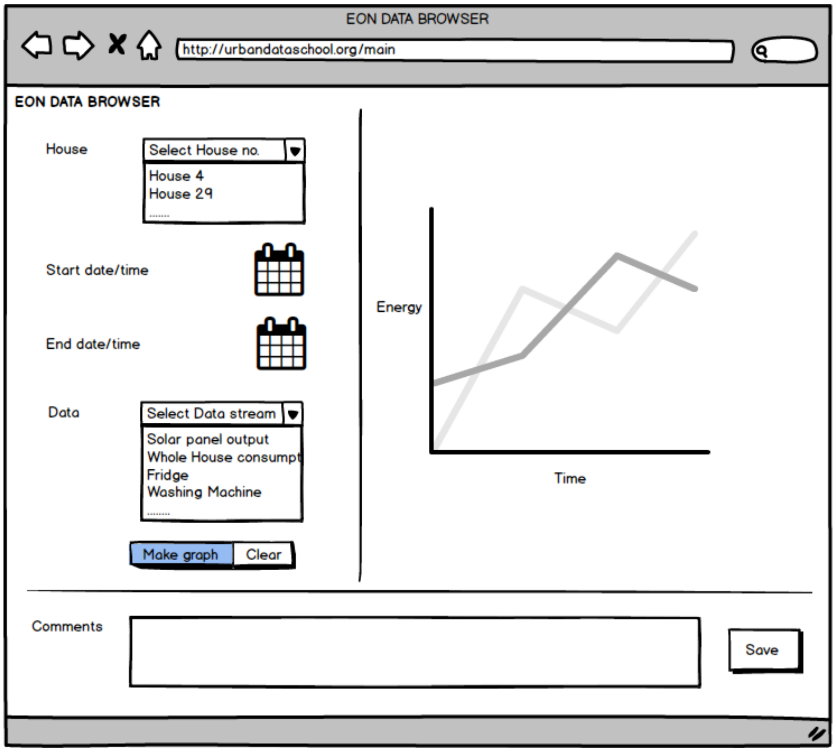

Figure 2: The mockup from which the Interactive Smart Meter Data Browser was created

5.2. Lesson Plan 2 (LP2): Potential for solar energy production

The aim of this lesson was to demonstrate, through data, that houses differ in their potential for producing solar energy, based on the direction they face and the size and pitch of their roof. This lesson used data that was derived from aerial photography, using LiDAR technology. This data set showed the potential energy production by installing solar panels on each building within the city. The data came from the local area and students were able to look at their school and their own houses (P3).

5.2.1. LP2 Activities

Students followed an inquiry process (P1) where they answered questions from the data. As in the smart meter example, the guided inquiry stage started with a representative snapshot of data (P2) from which they could see roughly the size of roofs and where a solar panel might go, colour coded according to whether it was predicted to give a low or high solar yield (figure

3). Students worked in groups on all activities. Students were prompted to answer the following questions:

- Which house is best for fitting solar panels to? Which is the worst?
- Look at the houses on the map, why do you think these are good/bad?

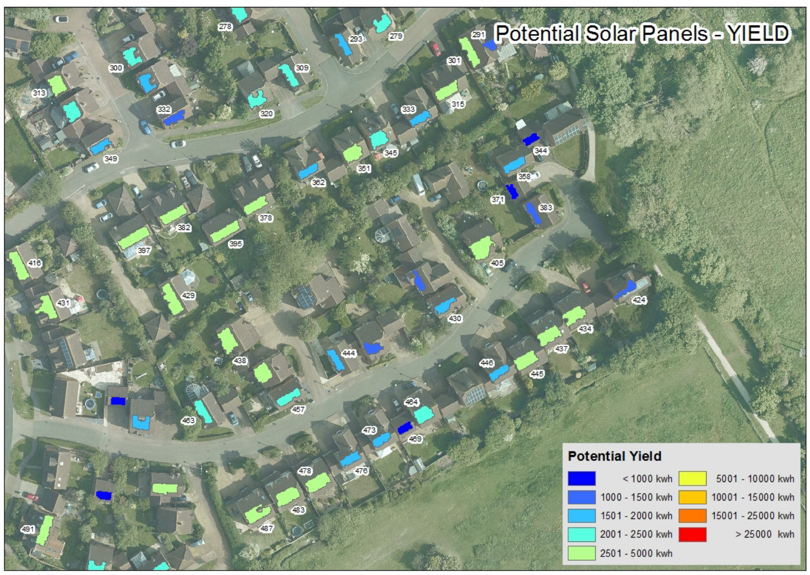

Figure 3: Solar potential data set

There were variations in how this stage was delivered. For LP2 there was no planned open inquiry stage as this was delivered only to younger students. Instead, the variations of the guided inquiry were:

No technology: Students were given a printout of the map and the snapshot area was an estate close to their school that they were all familiar with. The associated data could be found from a table from which they could look up each property by the ID and find data about the solar potential, orientation, size and pitch of roof as well as the estimated cost of the panel.

With technology: students used an interactive map that allowed them to zoom, pan, search by postcode, select the satellite or streetmap layer, and click on an area of the map to view data. This is shown in figure 4.

Through this, they could navigate across the city and ask and answer their own questions from the data, thus following the expansion principle (P2). In the guided inquiry stage, these students entered their own postcode to select a region of houses from their own area from which to answer the above questions.

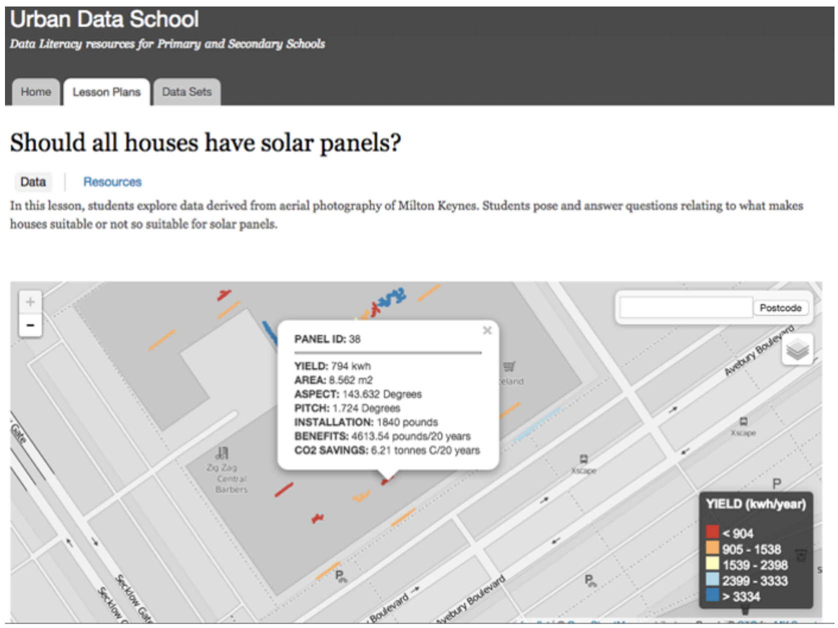

Figure 4: Urban Data School Solar Potential lesson plan showing the Interactive Solar Data Set

5.2.2. LP2 Outcomes

The intended outcomes were that students would: a) understand how roof size, pitch and direction affect solar yield; b) understand the difference between interpreting data from the map and from a table (e.g., ability to see things blocking solar panels compared to ability to do statistics); c) find errors in the data and understand that data can be flawed.

5.3. Lesson Plan 3 (LP3): Be a LiDAR device

The aim of this lesson was to provide students with the experience of capturing data by aerial survey. This activity is based on the personal data collection principle (P6).

5.3.1. LP3 Activities

Students were shown the principles of using light to measure distance, with the help of a portable laser measuring tool. Students then worked in groups and started by building their own house from plasticine onto which they marked a grid of 1cm by 1cm (figure 5). This follows the STEAM principle (P5). They then used home made rulers to measure the height of each square, transferring their data onto a sheet of paper. Groups then swapped their sheets, to see if they could understand the shape of the house from the data alone.

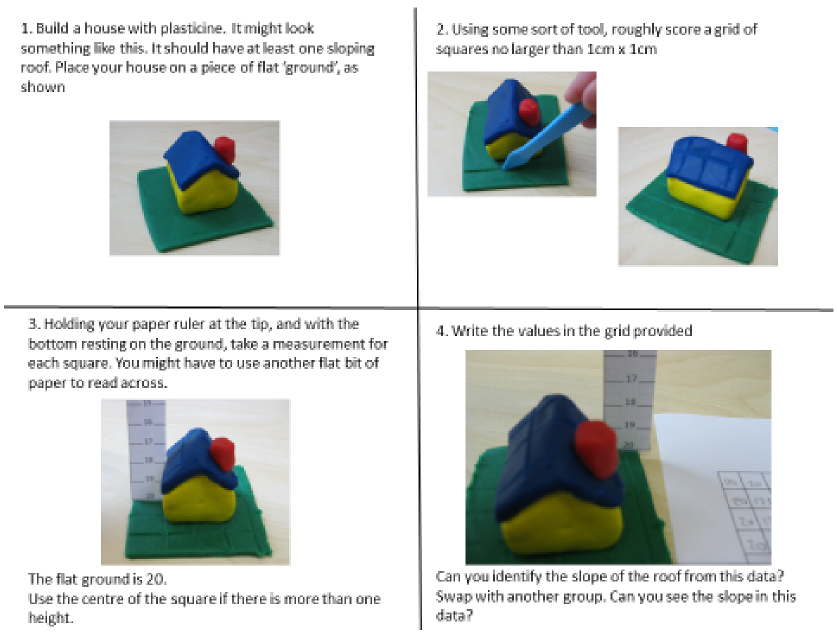

Figure 5: Steps for creating the plasticine house with grid

5.3.2. LP3 Outcomes

The intended outcomes were that students would understand how LiDAR data builds a picture of a landscape. They should also understand about data resolution and how this affects accuracy and the trade off between processing large data sets and having accurate measurements. A further aim was to improve their general understanding of how the data for the solar yield of roofs was created.

5.4. *Other activities*

We have described three lesson plans that were constructed and used across the field trials. We omit some activities that do not contribute to the later discussion and where, on the whole, the findings are reported elsewhere [29, 30]. One activity that should be mentioned is 'be your own smart meter', which encouraged students to collect their own energy data according to principle P6 and then to create novel visualisations from it. This was conducted each time in conjunction with LP1 to contextualise the smart meter data. The decision to exclude it was to reduce the amount of results to report - instead we have opted to discuss this principle in terms of LP2 and its complement LP3.

6. Methodology

We recruited three teachers to participate in four ethnographic field studies using the developed lesson materials within their classes. One teacher participated in two separate field studies in two different years of the project. Each field study comprised two or three classroom sessions in which students undertook the activities, usually at one week intervals. There was a constraint in recruiting schools, in that they needed to be in the geographic location covered by the data sets. Teachers were recruited through personal contact.

The constraints and method of recruitment meant that we ended up engaging with teachers of differing ages, subjects and abilities. Each field study was therefore adapted to align with the requirements of the teacher and their class. This process was led by the teachers, who were invited to select only activities that suited them and to adjust the design of these selected activities then decide how the teaching sessions would be delivered and who would lead: either the teacher, the researcher or a co-led session between teacher and researcher. In the classroom, all activities were undertaken by students in groups of 2 or more.

Evaluation at the end of each field study led to incremental improvements to the design and delivery of lesson plans, also taking into account the adaptations required by the teacher for the following field study. In addition, the technologies to support teaching were developed and used in the final two studies. This need for flexibility lent itself to a long-term qualitative approach to evaluation, rather than controlled studies where it would be possible to collect quantitative data.

6.1. Data collection and analysis

Data was collected for the purpose of refining the approach in a future iteration and also with a focus on assessing the students' ability to link questions to data and to start to form their own inquiries. Data was collected from students in both primary and secondary schools. The total age range of students participating in activities was between 10 and 14.

Each field study was observed by one or more Participant Observers (POs), who recorded videos or took photographs and made notes both during and after the sessions. Participant observation is useful for understanding how people relate, to each other and to task materials, and to identify future questions to be answered [31]. The observation procedures were discussed between observers beforehand. POs were tasked with noting when students needed help, in identifying parts of the lesson plans that caused problems and most importantly any evidence that students were thinking beyond the initial activities and posing their own questions from the data. POs were also tasked in noting down the number of students engaged in tasks and how they formed into groups. The level of participation of the observers varied from co-leading the session to supporting students in practical group work activities. As POs were busy during the sessions, the main data was captured in a summary that was written up as notes immediately after each session. Where practical, verbatim quotes of students were captured at the time, but this was not systematic.

At the end of each field study, the photographs, verbatim quotes and PO summaries were combined to create a single narrative about what was happening in the session, focusing on what problems were encountered and what questions did students ask.

In two field studies that were conducted with older, secondary school students, additional data was collected directly via worksheets and from classroom materials (such as post-it notes). This captured the questions that students asked from data at different points throughout the activities. A qualitative coding of this data to assess the questions for *answerability from the data* was undertaken by the first author, who had expertise with both the data and its use in research. It was verified by a second researcher, leading to some adjustments until a consensus was reached. This process aligns with the process undertaken by [21]. Both an inductive and deductive approach was taken to the coding. In this process, some initial categories were suggested and used to guide the first coding, then these were refined based on the analysis of each question.

Due to the longitudinal nature and slightly differing focus in each field study, the data collected was different in each case which made controlled experimentation difficult. However, each individual classroom session yielded rich data from observations and working materials.

7. Results

This section is structured according to the research questions listed in section 1. For clarity, results that do not contribute to this discussion will be reported on only minimally, or left out altogether.

There were four field studies; a total of 67 students took part. These are shown in Table 1 in the order in which they were conducted, approximately 6 months apart each time.

Id	Sessions	Year(age)	Pupils	Subject	Lead	POs	Activities
FS1	2	5 (10-11)	12	Maths	co-led	1	LP2 no tech
FS2	3	9 (13-14)	17	Triple science	teacher	2	LP1 no tech
FS3	2	7 (11-12)	25	Geography	researcher	1	LP1 with tech
FS4	2	5 (10-11)	13	Maths	co-led	2	LP2 with tech then

Table 1: Field Study details

Figure 6 summarises the findings from across the four field studies and details how they are used to answer the research questions. These findings are expanded upon in the remainder of the results section.

7.1. Answering RQ1: What factors influence students abilities to ask and answer questions from the presented data?

Lesson plan 1 was designed to follow standard inquiry processes (P1), starting with a guided inquiry and then moving to a more open inquiry with older students. Following the foundational competence principle (P4) and knowing that students may struggle in particular to relate questions and data - which is an important part of the inquiry process, especially an open inquiry - the following results explore the extent to which this was supported through the activities. The focus is on on a comparison between the FS2 and FS3 brainstorming activities of Lesson Plan 1 (see figure 7). This is the start of the open inquiry stage and it took place after all students had completed the guided inquiry from the snapshot of the data (first part of LP1). This

Research Question	Methods	Related Field studies	Participants: No. (age)	Activities	Analysis	Main findings
RQ1: What factors influence students' abilities to ask and answer questions from the presented data?	Comparison between 2 field studies with qualitative data collection through worksheets, classroom materials and participant observation	FS2 FS3	17 (13-14) 25 (11-12)	LP1 open inquiry stage, with smart meter data (both studies)	Categorization of questions framed by students in open inquiry	Younger students found it harder to ask questions directly from the data (see section 7.1)
RQ2: What is the role of data interaction in facilitating the inquiry process?	Comparison between 4 field studies with qualitative data collection through observation	FS1 FS2 FS3 FS4	12 (10-11) 17 (13-14) 25 (11-12) 13 (10-11)	LP2 (no tech) LP1 (no tech) LP1 (with tech) LP2 (with tech) Technology was in the form of an interactive data browser for a) smart meter data b) solar panel data	Narrative construction based on participant observation of students doing the activities	Students who were able to interact with the data were observed to start following their own inquiries, even when not prompted to by the worksheets (see section 7.2)
RQ3: How does personally collecting data changes one's perspective of it?	Single field study with qualitative data collection through observation	FS4	13 (10-11)	LP2 solar panel task, *followed by* LP3 LiDAR data collection task	Narrative construction based on observation and video recordings	Students become more critical of data when they gain experience in collecting it (see section 7.3)

Figure 6: Summary of results

relates to the categorisation of questions that students made in this stage (see row RQ1 of figure 6).

The question categories that were obtained through coding were as follows. We include also their alignment to the question categories used by Shelley et al. [21]. We have included the 'not answerable' category here, as this was originally suggested prior to coding taking place. However, this category was not needed in the end.

C1 Smart meter questions (completely answerable): students pose a question that can either be answered directly from a further analysis of the smart meter data, or where the further analysis could give enough information for them to form a reasonable hypothesis (that may then lead to further information being needed to verify).

Figure 7: Some students placing their brainstorming questions onto a whiteboard

C2 Supplementary questions (conditionally answerable): students would require further data or information to answer the question, but this answer would help to interpret findings from the smart meter data.

C3 Topic questions: questions students have that aid general understanding of the topic, but are not directly related to the smart meter data.

C4 Validity questions: students query the validity of the data.

C5 Not answerable: the question is out of scope for both the topic and the data.

It should be noted that, in categorising questions, the goal was to assess the ability of the students to frame questions around the smart meter data for which they could offer a line of reasoning by which their proposed analysis may provide an answer to their question, rather than to judge the quality of this reasoning. Hence, the first category combined questions that could be answered from the smart meter data and those for which the analysis could lead them to form a hypothesis that might then need verification from additional data. Therefore, in completing the categorisation, attention was paid to the explanations given by the students either in their workbooks or in discussion with the teacher or researcher (which were recorded as observations). Where students could offer a plausible explanation of what they would be looking for from the data and how this would relate to the question, the question was placed into the first category. To give an example, one teacher queried how students would tell from data if there were a young family in the house. A student offered an explanation that the "mini-spike in the energy data could indicate a young family having to heat food, put music on". With regard to the possibility of visitors being in the house between 8:00 and 12:00, a student suggested they "could check whether this happens every day by looking for a spike on other days".

Next, we counted the questions that appeared in each category. We did this separately for FS2 and FS3, to enable comparisons between them. Figure 9 lists all of the questions in the FS2 session and how they were categorised. Additionally, we know whether these questions were selected for further analysis in the refinement stage and by how many students. This information is also presented in the table (it will be discussed in more detail in the next section). It should be noted that some students did not specify in their workbooks which questions they had selected for the further analysis, whereas some students decided to write down new questions that had not been presented by the whole class in the brainstorming stage.

FS2 students posed a total of 18 questions across the two stages (brain-

	Brainstorm (shaded questions added at refinement stage)	selected
C1	Does the house have the same pattern every day? we would need another 6 more graphs to compare	2
	Do you have a young family?	3
	Were there visitors in the house between 8am-12pm?	1
	Is it a house full of adults or a family?	2
	What is the average amount of energy consumption of that day shown in the graph? (per hour) (House 4)	0
	If there's no children, do the adults who live there work on a schedule (e.g. 9-5) or work irregular shifts?	3
	Did the whole street's power go, not just that single house?	7
	Is there the same amount of energy used on weekends and weekdays? Do different households use the same amount of energy? Is there always power cuts every once in a while? If answer yes or no – please explain why?	2
	Does this house have the same energy consumption to other houses?	1
	Are the adults employed or unemployed?	1
	What season is it?	1
	How much energy is used at this busy time of year?	1
	When does the electricity usage nationally peak? I chose this question because you can see when a household is most active.	1
C2	Who might be living in House 4? And how many?	2
	If family home, is child(ren) home-schooled?	0
	Where is the location (e.g. countryside or city)?	0
	What is the day? A weekday or weekend? (*note, the smart meter data only indicates dates, not weekdays)	1
C3	What is your smart meter data?	0
C4	- None asked	

Figure 8: Questions asked in FS2 related to energy consumption

storm and refinement). The majority of questions in the refinement stage were chosen from those where the answer was in the data (25) compared to from additional data (3) or general topic (none) indicating that their understanding of how to select good questions was improving through the class discussions and use of technology to interact with the data.

Figure 10 shows the questions asked by students in FS3. They did not formally write down questions for the refinement stage, so this information is missing from the table, but is discussed (based on the observations) in the next section 7.3. FS3 students asked a similar number of questions as FS2, despite a greater number of students (25 students compared to 17). In both field studies, the students worked in groups of two or three.

As described in row RQ1 of figure 6, the notable result is that FS3 students asked fewer questions of the data and more about the data, indicating some difficulty in framing these types of questions. For example, FS3 stu-

	Brainstorm
C1	How often do power cuts happen?
C2	How much energy does the average family use?
	What household item uses the most energy?
	How is it possible to not use electricity in a day on the weekends?
	Why do the family have no electricity in the middle of the night?
	Why do they use less during the middle of the day?
C3	Would energy ever run out?
	How much money averagely spent on energy?
	Why do you have to use a smart meter?
	What happens when too much energy is formed, does the smart meter warn them?
	How much is the smart meter?
	How do smart meters measure microwaves' or toasters' energy use?
	Where else other than homes do you get smart meters?
C4	Does the smart meter always collect energy every day, hour and second. Does it ever stop working?

Figure 9: Questions asked in FS3 related to energy consumption

dents noticed that less energy was being used in the middle of the day and asked why. On the other hand, FS2 students framed much more specific questions that could be answered by looking at more data from the smart meter data set, such as "Does the house have the same pattern every day? We would need another six more graphs to compare." FS3 students also had many more questions that would aid their general understanding of the topic (C3). The differences between FS2 and FS3 were the age of students (FS3 students were approximately 2 years younger) and the lesson's subject (science in FS2, geography in FS3).

Overall, the students were able to:

- frame new questions of the wider data set after initially focusing on just a very small part of it;
- create plausible explanations of their findings - even if sometimes the explanations were not the only possible ones and even though they were often not verifiable without additional information.

7.2. Answering RQ2: what is the role of data interaction in facilitating the inquiry process?

This section compares the lesson plans, LP1 and LP2, undertaken firstly without technology and secondly with the use of an interactive data browser

- in each case, by a different set of students at a different point in time.

7.2.1. Technology use in FS2 and FS3

FS2, in which LP1 was conducted without the use of technology, is described in the previous section 7.1. This section focuses on the refinement stage in FS3, in which students were able to ask and answer questions rapidly using an interactive tool in which they could select the smart meter energy consumption data for a time period and a house in which they were interested (see figure 6, RQ2: comparing LP1 with and without technology). They could also view data at the appliance level. This data came from smart plugs, which could be configured by each individual household.

The data in this stage is based on the observations of the participant observers (POs), as these students did not have time to write their findings in the book. Observations were based on what students were looking at and on summaries of the conversations that students in a group had with each other, or with the PO. Any interpretations presented in these results are based on the interpretations made and written by the POs at the time.

The observers noted that students could quickly grasp the meaning of the graph without any help at all, and were starting to answer questions immediately about the times of highest/lowest energy use, as well as starting to propose theories for what caused them (see the findings for RQ2 in figure 6). Students could also easily identify the relationship between the graph and daily life activities of the occupants of the houses. This was evidenced through the stories that students told about what they thought was happening in the house, based on the data. In this case, students tended to focus on questions that compared either a single property or appliance across different time periods. One explanation for this is that the interactive tool made selection of appliances and time periods easier than changing to view a different property. Although it is not clear from the mock-up in figure 2, there was one additional button to press to select the data set of a different house.

The queries and explanations were analysed and categorised using the same process as for the questions (section 7.1). These questions (by nature of the task) all belonged in category *C1*, in that they were *completely answerable* from the data so the aim was to undertake a deeper analysis of the types of questions that fell within this category to show what students were most interested in. This analysis revealed that questions fell broadly into two categories. These are now discussed, with some representative examples of explanations.

Comparing a single property at different times: One group found a reduction in energy consumption at Christmas hypothesising that the family may have spent Christmas elsewhere. Another group focused on anomalies, first discussing possible reasons for a zero value, including the possibility of a power cut. Another student in the group said a power cut would last longer, so perhaps a fuse had gone in the house and the person had woken up and gone and flicked the fuse box back very quickly. Another student thought that perhaps it was a key meter. This same group also noticed two spikes in the data, which they discussed with the researcher, leading to the explanation that perhaps the smart meter was in error.

Comparing a single appliance at different times: One group was looking at TV consumption and found that the family had suddenly stopped using the TV. They speculated that the TV was broken, but could not think of any other reason, for example, they did not know that the smart plug might have been moved and used to monitor something else. When told this, they decided that this was a more likely explanation.

7.2.2. Technology use in FS1 and FS4

In FS1 students undertook activities related to LP2 (solar potential) using paper-based maps and associated data sets given in a printed table (see RQ2 figure 6). The aim was that students would understand how direction, roof area and pitch contributed to solar yield. Students worked in groups. At the end they presented their findings. Their conclusions after engaging with the task were:

- "If the house [roofs] are slanted then they have the most chance of getting the most electricity."
- A 4-sided roof would be "harder to put solar panels on, because sometimes the sun doesn't come from that side."
- "It's best if the solar panels are facing south, because that's the direction of the sun in the day."
- "Even if you buy these really big expensive solar panels, it might not make much of a difference - it might be a waste of money."

Overall, these answers reveal that students had picked up important principles about solar panels through interpreting the dataset. These include that

Figure 10: Interacting with the solar map

the roofs must face a certain direction and be slanted to get the most sun. They had also begun to understand some of the cost implications.

In FS4, students followed the same set of activities, but they used an interactive version of a map showing solar potential of all roofs in the city (see RQ2 figure 6). Students, working in groups of two or three, first undertook the guided inquiry stage based on putting in their own postcodes (Figure 9). This normally revealed an area of about 20 houses.

The following data is based on the observations made by the POs at the time. Students were observed to start asking further questions independently very quickly and navigating the map to try to find the answer (see findings for RQ2 in figure 6). For example, one group very quickly put in the postcode for their school. They discovered an anomaly in the data, where a non-building in the school was identified as having good potential for fitting a solar panel. Another group tried to find a building (a head office of a famous pizza chain) that they knew "has a very big roof" to see how much the panels would cost and how much energy it would produce. By querying the data more closely, students latched onto the idea of cost/benefit trade-off. This was despite such activities not being prompted: these students were meant to still be in a *guided* inquiry and there were no open inquiry activities planned for these

Figure 11: Measuring the slope of the roof.

younger students.

7.3. *Answering RQ3: How does personally collecting data changes one's perspective of it?*

The following results focus on the LP3 activities of FS4, which took place directly after the LP2 activities described above, where students were exploring the LiDAR data through the technology. The data is based on analysing and constructing a narrative from the observations of the POs and video data from the session.

There were three groups completing the task, with 2 or 3 children in each group. All groups completed the task of creating and measuring the house (figure 11).

One observed group were able to complete their grid of height values taken by measuring the roof height for each square they had drawn onto the house and then begin to identify the slope of the roof from the data alone. With some support from the PO, they were working out how they would tell just from data which way the house was facing (figure 12).

Two of three groups swapped their grids and were able to find the slope from the other group's data, with one group correctly identifying that the

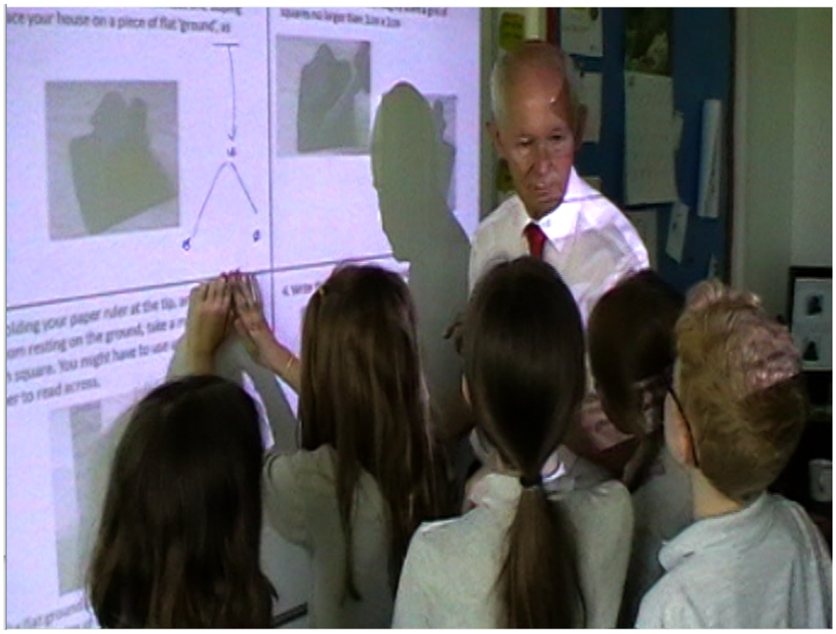

Figure 12: Recreating the house from data

other group had made a house with a ‘wiggly roof’ and then asking to see the house for themselves.

At the end of this classroom session, there was a general discussion. The noted observations were as follows:

- Students commented how “stupid” the data is, because it “doesn’t know it is looking at a house, or someone’s back garden”.
- Students could easily think of things that might have slopes that the aerial survey might pick up but were not roofs, including bus shelters or hills.
- Students thought that it was normally better for humans to process visual data, but when the data set is so large (as in this case), then it is good to give some intelligence to computers so they can help.
- This last comment prompted a discussion about how to add more intelligence to the data processing algorithm. One suggestion was that bus shelters would not have such a steep pitch. Students started thinking about combining data sets, proposing that one way to tell a house from

other buildings through the data was to measure the heat of people inside.

Taking all of the above into consideration, it appears that the LiDAR task has prompted a good level of understanding of the potential and limitations of the data set (see main findings for RQ3 in figure 6), whilst the initial task with the interactive map prompted more free exploration and asking questions from the data itself.

8. Discussion

We begin this discussion by considering what has been found with regard to students' abilities to ask and answer questions from externally sourced data.

In the fourth field study (FS4), when students were able to directly interact with data through the data browser (figure 3), they became very keen to start driving their own inquiries, even though this was not an explicit part of the task (RQ2 in figure 6). For example, deciding to look at the cost of solar panels on a large roof and finding out whether their own houses should get solar panels or not. As pointed out by Konold and Higgins [22], data investigations start with questions about the real world - but such questions must be revised to ones that can be answered from data. The expansion principle (P2) was proposed as a way to support this, by engaging students first with a data snapshot and then allowing them to navigate across the wider data set.

In this regard, the finding of note was that younger students (FS3) had more difficulty than older students (FS2) in framing inquiry questions directly from data, when engaging with only a single snapshot (RQ1 in figure 6). Older students were more likely to choose questions for which they could present a plausible explanation of what they would look for in the data to answer. Both sets of students had undertaken an identical task, so the main factor on which to understand the difference was their age. This supports findings of [32] that students of this age find it difficult to link questions, data and explanations coherently. If we take the perspective of Piaget [33], the younger student group are just at the start of their formal operational stage, where they gain the ability to reason in abstract forms. Prior to this stage, children have more reliance on concrete manipulation. If this is the case, then it could explain the observations that the younger students asked more focused questions when they used the technology to engage with the data. However, data collected regarding the role of technology was too sparse to be able to draw firm conclusions and future work would need to investigate more thoroughly the extent to which the technology supported this adaptation of question strategy and played a role in supporting the expansion principle.

Turning attention to the personal data collection principle (P6), it was notable that students in the smart meter task (LP1) consistently proposed a supply failure as the reason for a zero reading, whereas a more plausible explanation given the very brief time of the zero reading was that the meter itself had failed. While the results reported have been quite focused, it is fair to mention here that these tasks were conducted across a two year period in a number of settings. It was observed across a number of engagements with smart meter data and also the solar panel data set that students were reluctant to attribute errors to the measuring instrument.

In previous work by Hautea et al. [12] it was discovered that young people became sceptical about data through their interactions with it. In this set-up, the students (of a similar age range to the ones in these studies) were interacting with data in an environment in which they were also contributing to the data, so in effect the personal data collection principles was in place to help the students to understand better the possible source of errors. Similarly, in our studies when students started to collect data and became a LiDAR measuring instrument, they were more critical of the data (RQ3 in figure 6). These same students had interacted with the LiDAR-obtained solar data in the previous week and had been observed to focus on driving their own inquiries from the data to find if houses were more or less suitable for solar panels. However, in the following week when they were learning how the data was collected, they began questioning whether every 'roof' picked up in the dataset was a viable building for fitting solar panels and even started to think of ways to refine the processing of data to reduce such errors. This seems to support the **personal data collection principle** (P6), that students should collect data themselves to help them to interpret data and that this process of interacting with familiar data may be important in fostering data scepticism. In this regard, it would have been better to have these activities occur in the alternate order, so that students would first understand how the data was collected and then explore the data set.

The personal data collection principle should be investigated in a more controlled manner, to really understand the relationship between familiarity with data and ability to critique. It has wide-ranging implications for people's ability to use externally sourced data, whether it is for business needs, for empowerment or for innovation from data.

Finally, this work has demonstrated the many different ways that these types of less typical classroom data and smart city concepts can be integrated in a school curriculum and how activities can be designed around them in a way to support development of critical data literacies. Overall, the lesson plans can be shown to achieve their intended outcomes. In the first lesson plan, students showed evidence of finding and explaining common patterns in energy data. In the second lesson plan, students demonstrated a good understanding of the different factors that effect solar yield. In the third lesson plan, students came to understand how to recreate the 3D world from 2D data and the possible sources of error that came from the measuring technique. However, this was not the end of the story. Students showed evidence of learning a lot more, for example about the domain of energy, the importance of being energy efficient and the pros and cons of solar energy as a renewable source.

9. Conclusions

This paper presents findings from an initiative to take complex data from a smart city project into schools and to use it as a teaching resource. It explores the use of data literacy activity design principles to support the co-creation, with teachers, of the teaching resources and the development of technology to support interaction with data. The project followed a research through design approach which created an initial set of teaching materials that were refined each time they were taken to a new classroom and also adapted by the teacher to fit the new context. The technologies to support data interaction were designed to have limited functionality and to support just a small part of the classroom delivery, which also included workbook activities, and practical tasks.

The main findings were that:

- younger students require support in framing inquiry questions that can be answered from externally sourced data;

- when engaging with externally sourced data it can be useful to act in the role of a data collector to understand better where errors can creep into the data and to develop better data scepticism.

Overall, the learning of data skills lends itself very well to cross-curricular learning and can begin with students as young as ten years old, as evidenced through the variety of school contexts in which we worked. Data literacy activity design principles provide a way to structure learning from external data sets. This may support teachers to develop new activities from open data. The teaching of data in context is important and local, open data can be a good resource for teaching, if supported in the right way.

10. Acknowledgements

This work was undertaken as part of MK:Smart, supported by the Higher Education Funding Council for England [Q-13-037-EM, 201417]. We would like to thank the teachers who welcomed us into their schools and who made this work possible.

11. References

[1] J. Mateos-Garcia, G. Windsor, S. Roseveare, Analytic britain: Securing the right skills for the data-driven economy, London: Nesta (2015).

[2] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, J. Crowcroft, Human-data interaction: The human face of the data-driven society (2014).

[3] R. Bhargava, E. Deahl, E. Letouzé, A. Noonan, D. Sangokoya, N. Shoup, Beyond data literacy: reinventing community engagement and empowerment in the age of data, Data-Pop Alliance White Paper Series. Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Lad and Overseas Development Institute) and Internews (2015).

[4] P. Cobb, K. McClain, Guiding inquiry-based math learning, Cambridge University Press, 2006.

[5] M. Janssen, Y. Charalabidis, A. Zuiderwijk, Benefits, adoption barriers and myths of open data and open government, Information systems management 29 (2012) 258–268.

[6] K. Kastens, M. Turrin, Geoscience data puzzles: developing students' ability to make meaning from data, in: AGU Fall Meeting Abstracts.

[7] L. Bowler, A. Acker, W. Jeng, Y. Chi, "it lives all around us": Aspects of data literacy in teen's lives, Proceedings of the Association for Information Science and Technology 54 (2017) 27–35.

[8] J. Zimmerman, J. Forlizzi, S. Evenson, Research through design as a method for interaction design research in hci, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp. 493–502.

[9] W. Gaver, What should we expect from research through design?, in: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp. 937–946.

[10] A. Wolff, D. Gooch, J. J. Cavero Montaner, U. Rashid, G. Kortuem, Creating an understanding of data literacy for a data-driven society, Journal of Community Informatics 12 (2017) In–press.

[11] E. Deahl, Better the data you know: Developing youth data literacy in schools and informal learning environments (2014).

[12] S. Hautea, S. Dasgupta, B. M. Hill, Youth perspectives on critical data literacies, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp. 919–930.

[13] B. Alper, N. H. Riche, F. Chevalier, J. Boy, M. Sezgin, Visualization literacy at elementary school, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp. 5485–5497.

[14] A. V. Maltese, J. A. Harsh, D. Svetina, Data visualization literacy: Investigating data interpretation along the noviceexpert continuum, Journal of College Science Teaching 45 (2015) 84–90.

[15] P. Wiegand, Children's understanding of maps, International Research in Geographical and Environmental Education 8 (1999) 66–68.

[16] S. Srikant, V. Aggarwal, Introducing data science to school kids, in: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, ACM, pp. 561–566.

[17] V. R. Lee, J. R. Drake, J. L. Thayne, Appropriating quantified self technologies to support elementary statistical teaching and learning, IEEE Transactions on Learning Technologies 9 (2016) 354–365.

[18] S. Van Wart, T. S. Parikh, Increasing youth and community agency in gis, in: GeoHCI Workshop at CHI.

[19] S. Van Wart, K. J. Tsai, T. Parikh, Local ground: A paper-based toolkit for documenting local geo-spatial knowledge, in: Proceedings of the First ACM Symposium on Computing for Development, ACM, p. 11.

[20] S. Dasgupta, B. M. Hill, Scratch community blocks: Supporting children as data scientists, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp. 3620–3631.

[21] T. R. Shelley, C. Dasgupta, A. Silva, L. Lyons, T. Moher, Photomat: A mobile tool for aiding in student construction of research questions and data analysis, Technology, Knowledge and Learning 20 (2015) 85–92.

[22] C. Konold, T. L. Higgins, Reasoning about data, A research companion to principles and standards for school mathematics 193215 (2003).

[23] S. Williams, E. Deahl, L. Rubel, V. Lim, City digits: Developing socially-grounded data literacy using digital tools, Journal of Digital Media Literacy (2014).

[24] R. Bhargava, R. Kadouaki, E. Bhargava, G. Castro, C. D'Ignazio, Data murals: Using the arts to build data literacy, The Journal of Community Informatics 12 (2016).

[25] C. Anslow, J. Brosz, F. Maurer, M. Boyes, Datathons: an experience report of data hackathons for data science education, in: Proceedings of the 47th ACM Technical Symposium on Computing Science Education, ACM, pp. 615–620.

[26] C. D'Ignazio, Creative data literacy, Information Design Journal 23 (2017) 6–18.

[27] C. Maybee, L. Zilinski, Data informed learning: A next phase data literacy framework for higher education, in: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, American Society for Information Science, p. 108.

[28] R. Bhargava, C. D'Ignazio, Designing tools and activities for data literacy learners, in: Workshop on Data Literacy, Webscience.

[29] A. Wolff, A.-M. Valdez, M. Barker, S. Potter, D. Gooch, E. Giles, J. Miles, Engaging with the smart city through urban data games, in: Playable Cities, Springer, 2017, pp. 47–66.

[30] A. Wolff, J. J. C. Montaner, G. Kortuem, Urban data in the primary classroom: bringing data literacy to the uk curriculum, The Journal of Community Informatics 12 (2016).

[31] B. B. Kawulich, Participant observation as a data collection method, in: Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, volume 6.

[32] H.-K. Wu, C.-E. Hsieh, Developing sixth graders inquiry skills to construct explanations in inquiry-based learning environments, International Journal of Science Education 28 (2006) 1289–1313.

[33] J. Piaget, B. Inhelder, The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures, Routledge, 2013.